Center and Spread: A Pas de Deux

Clifford Konold
Scientific Reasoning Research Institute
University of Massachusetts, Amherst

Alexander Pollatsek
Department of Psychology
University of Massachusetts, Amherst

Paper presented as part of the Symposium: There's More to Life Than Centers
Research Presession, 77th Annual Meeting of NCTM
San Francisco, CA. April, 1999

In *Full House*, Stephen Gould (1996) argued that

> we are still suffering from a legacy as old as Plato, a tendency to abstract a single ideal of average as the "essence" of a system, and to devalue or ignore variation among the individuals that constitute the full population (p. 40).

In warning of the dangers of a mindless reliance on averages, Gould claimed that variation is a "concrete reality" while averages are "mental abstractions," adding that "no actual individual can stand for the category's deeper reality" (p. 41). These sentiments echo the concern that prompted this symposium, namely that statistics educators and researchers have been emphasizing measures of center to the detriment of measures of spread.

Gould's analysis certainly applies to the practices of social scientists, many of whom are fixated on measures of center. Over the years, many professional organizations and their journal editors have tried with limited success to convince their members of the value of reporting statistical measures that are more informative about sample variability than the ubiquitous means and p values. For example, a special task force commissioned by the American Psychological Association (1996) recommended that journal articles include:

> (a) more extensive descriptions of the data.... This should include means, standard deviations, sample sizes, five-point summaries, box-and-whisker plots, other graphics, and descriptions related to missing data as appropriate.

> (b) enhanced characterization of the results of analyses (beyond simple p value statements) to include both direction and size of effect ..., and their confidence intervals....

Changing the policy on which statistics get reported in journal articles is not likely to coax the user of statistics out from under the rock " average"; what we need is a change of thinking. One can hope that current efforts to teach statistics to younger students will help effect such a change (see Lajoie, 1998).

In thinking about what our younger students need, we might conclude from Gould (1996), and from the statistical tunnel-vision of social scientists, that elementary and secondary students are also prone to fixate on measures such as means and medians and that we should therefore administer larger doses of standard deviations and IQRs. However, we believe this approach would be a mistake. For although Gould's claim accurately describes tendencies of some of those with considerable training in statistics, it does not describe how students with no or moderate training in statistics reason about data.

In this article we argue that the focus on centers of distributions in current statistics instruction isn't too excessive, but rather of the wrong kind. Exploration of centers ought to be seen as part of a study of characteristics of complex, variable processes; too frequently, centers are portrayed as little more than summaries of groups of values. To highlight this difference, we examine how statisticians use and think about measures of center to compare two groups, and contrast this with what researchers have observed students doing. We also present various commonly-held interpretations of averages and show how most of these interpretations provide little or no conceptual basis for comparing groups. Based on our analyses, we offer several recommendations about how to help students come to see measures of center and spread as co-constructed ideas.

A Statistical Perspective

Every four years the Federal Government conducts the National Assessment of Educational Progress (NAEP), an assessment of student capabilities in grades 4, 8 and 12. The most recent results from the reading assessment, which were released in the spring of 1999, were interpreted as indicating an increase of reading scores (Donahue, Voelkl, Campbell, & Mazzeo, 1999). The average score for the 8th graders was 264 (out of 500 possible points and a standard deviation of 50). This average was up from 260 where it had been in 1994. The twelfth graders had also gained four points on average since 1994, and the fourth graders 3 points.

Reports such as this are now so commonplace that we seldom question the logic of this reasoning. What is the rationale in this case for comparing group means and, having done so, for taking seriously the apparently small difference between those means? To answer these questions from a statistical perspective requires describing the idea of a central tendency.

Central Tendencies

A statistician sees the mean of each group as a stable property of a variable system, a property that becomes evident only in the aggregate. This stability can be thought of as the certainty in situations involving uncertainty, the signal in noisy processes, or, the descriptor we use here, central tendency. By *central tendency* we refer to (a) a stable property of a variable process that (b) is embodied by a value derived from the output of that process and (c) is better approximated as the number of observations considered grows. The obvious examples of statistics that could be used to approximate a central tendency are means and medians. The notion of the mean as central tendency has two fundamental components: a stable characteristic, which is summarized by the mean, and a variable component, which can be summarized by a quantity such as the standard deviation. Thus, viewing a mean as a central tendency of a particular process requires acknowledging the fundamental variability inherent in that process. Because of this, we claim that the notion of a center understood as a central tendency is inseparable from the notion of spread.

In the above description, we could have spoken of populations rather than processes, but we have come to prefer the later term. One reason is that the notion of process better covers the range of statistical situations, many of which have no real population (e.g., weighing an object repeatedly). Another is that when statisticians begin thinking about why two samples might differ, they typically consider factors that may have played a role in producing the data, and thus about the underlying processes rather than static characteristics of the populations from which data were taken (not always by formal sampling methods). We further develop the NAEP example below to illustrate this type of reasoning about process. Frick (in press) argues that the difference between processes and populations is more than a semantic one, claiming that the tension between theoretical descriptions of random sampling and actual practice could be resolved if statisticians thought explicitly of sampling from processes rather than from populations.

Interpreted as a central tendency, the mean of 264 is a measure of a complex process that determines how well children in the U.S. read. An obvious component of this process is the reading instruction children receive in school. Another influence is the behavior of adults in the home: their reading habits, time spent reading to their children, the kind and quantity of reading material in the home. Reading skills are also affected by factors operating outside the home and school, including determinants of public health and development such as nutrition levels and availability and use of prenatal care, genetic factors, and the value placed on literacy and education by local communities and the society at large.

In taking a statistical perspective, we regard all of these influences together (along with many others including those of which we might be unaware) as a process that turns out readers of different capabilities. Even though readers produced by this complex process vary in their performance, we can regard the entire process at any given point in time as having a certain stable capability to produce proficient readers. The mean performance of a large sample of readers produced by this process is one way to measure the process' power (or propensity) to produce a literate citizenry. As Mme de Stäel explained in 1820, "events which depend on a multitude of diverse combinations have a periodic recurrence, a fixed proportion, when the observations result from a large

number of chances" (as quoted in Hacking, 1990, pg. 41.) And because of the convergence property of central tendencies, the larger the sample of readers, the better estimate we expect the sample average to be of the stable component of the process.

Given the huge sample size in the NAEP study (about 11,000 8th graders) and assuming proper care in composing the sample, we expect that the sample mean of 264 is very close to this propensity. Assuming that the 1994 mean is of equal quality, we can be fairly certain that the difference between these two means reflects a real change for the better in the underlying process that affects reading scores. Of course, the change of the mean between these years may not be meaningful in the context of literacy if one views the main goal to ensure that all students attain a minimal competence. However, even in that case we can abstract a single tendency out of the group data: the percentage of students in the sample reaching or exceeding some minimal score. According to the NAEP report (Donahue et al., 1999), the percentage of 8th graders performing at or above a basic reading level increased from 70% in 1994 to 74% in 1998. This percentage is formally a mean[1], and changes in it are also revealing of changes in the underlying educational process.

As long as a process remains stable, we expect the average output from that process to remain unchanged from sample to sample. Conversely, when the average of large samples changes, we assume that the process has changed in some way. We rely on these expectations to test efforts to alter processes. In the case of reading, we might introduce new curricula, run an ad campaign encouraging parents to read to their children, expand the school free lunch program in disadvantaged areas, upgrade local libraries. If we do one or more of these things and the average reading scores of an appropriate sample of children increases, we conclude that we have changed for the better the process for producing readers. It would be impossible to determine the effectiveness of these efforts by looking closely at individual cases.

---

[1] If we code people who perform at threshold or better as 1 and the others as 0, then the proportion of people who exceed threshold is the mean of the 1s and 0s.

Early Applications of Central Tendency

It was Tycho Brache in the late 1500s who introduced the use of means as central tendencies to astronomy (Plackett, 1970). He used them in addressing a problem that had long troubled astronomers — what to take as the position of a star given that the observed coordinates at a given time tended to vary from observation to observation. Taking the mean of multiple observations became the standard solution only after it had been determined that the mean tended to stabilize on a particular value as the number of observations increased. It was this idea of the mean as a central tendency that Quetelet in the 1800s began applying to social and human phenomenon (Quetelet, 1842). The idea of applying means to such situations was inspired partly by the observation that national rates of birth, marriage, suicides — events that at one level were subject to human choice — remained relatively stable from year to year. Some, including Arbuthnot and De Moivre, had taken these stable rates as evidence of supernatural design. Quetelet explained them by seeing collections of individual behaviors or events as analogous to repeated observations. Thus, he regarded observing the weights of 1000 different men, weights which varied from man to man, as analogous to, say, weighing the same man 1000 times, weights in this case varying from trial to trial.

How Novices Compare Groups

Having reviewed how statisticians interpret means and use them to make inferences about underlying processes, we now look at how non-experts compare groups and interpret measures of center. Group comparison is a rich context in which to explore the nature of peoples' thinking. It is the paradigmatic case in statistics. It also has face validity in that a problem such as deciding which of two groups is better provides a compelling reason for computing and using a statistic such as a mean or median. Contrast this with textbook situations in which students are asked to "summarize" data. In such cases it is often unclear why one would want to resort to simple descriptors when all the data can be displayed. We look at a few of these examples later in the paper.

Novices Tend Not to Use Averages

A number of researchers have reported that comparing two groups is a difficult task for students. Gal, Rothchild and Wagner (1990) interviewed students in grades 3, 6 and 9 to determine, among other things, their understanding of how means were computed and what they were useful for. They also gave the students nine pairs of distributions and asked them to decided for each pair whether the two groups were different or not. Only half of the 6th and 9th grade students who knew how to compute means (and, who could also, to some extent, interpret them) went on to use means to compare two groups, even when the groups were of unequal size. Similar findings have been reported by Hancock, Kaput and Goldsmith (1992) and more recently by Watson and Moritz (1999).

This difficulty is not limited to the use of means. Bright and Friel (1998) questioned 8th grade students about a stem-and leaf plot that showed the heights of 28 students who did not play basketball. They then showed them a stem-and-leaf plot that included this data along with the heights of 23 basketball players. This latter plot is shown in Figure 1 (from Bright & Friel, 1998, p. 81). Heights of basketball players were indicated as they are here in bold type[2].

Figure 1. Heights of Students and Basketball Players (**bold**)

10	
11	
12	
13	8 8 8 9
14	1 2 4 7 7 7
15	0 0 1 1 1 1 2 2 2 2 3 3 5 6 6 7 8
16	
17	1
18	**0 3 5**
19	**0 2 5 7 8 8**
20	**0 0 2 3 5 5 5 5 7**
21	**0 0 0 5**
22	**0**
23	

[2] In this plot, the row headed by 13 (the stem) contains four cases (leaves), three students each of 138 centimeters, and a fourth student of 139 centimeters.

Asked about the "typical height" in the single distribution of the non-basketball players, the students responded by specifying ranges within the distribution (e.g., 150-160 cm), a seemingly reasonable group summary. But shown the plot with both distributions, they could not generalize this method or find another way to determine "How much taller are basketball players than students?" In the words of Bright and Friel (1998, p. 80), these students could

> describe a 'typical' student or basketball player, but they did not make the inference that the 'typical difference' in heights could be represented by the 'difference in typicals.'

Surprisingly, we found similar difficulties among high school seniors who had just completed a yearlong course in probability and statistics (Konold, Pollatsek, Well, & Gagnon, 1997). On many occasions during the course they had used medians, primarily in context of box plot displays, as well as means to make comparisons among groups. But in the classroom they were supported by the curricula, software, and instructor, and thus to a large extent were not choosing for themselves the methods they would use. During a post-course interview they were less constrained in their choice of methods, and here they seldom used medians, means, or percentages when comparing two groups.

Comparing Frequencies of Similar Values in Two Groups

A method we see many students use to compare groups involves comparing actual numbers of cases within a narrow range (or "slice") of the dependent variable. For example, one pair of students we interviewed wanted to determine whether students with curfews studied more hours per week than students without curfews (Konold et al, 1997). They were familiar with both the data set they were investigating and the data analysis software they were using. The data set included a variety of information obtained from 154 students at their school. To explore this question during the interview, the students generated the two-way frequency table shown in Table 1. (To fit this table on the page, we have omitted columns as indicated by the dotted lines.)

Table 1. Homework hours of students with and without curfews

Curfew	Hw 0	12	14	15	27	total
no	7 (0.14)	1 (0.02)	2 (0.04)	4 (0.08)	1 (0.02)	50
yes	2 (0.02)	3 (0.03)	5 (0.05)	5 (0.05)	0 (0.00)	100
total	9 (0.06)	4 (0.03)	7 (0.05)	9 (0.06)	1 (0.01)	150

After the table appeared, one student asked the other:

S1: What was your question again?

S2: If having a curfew affects your studying, like you study more if you have a curfew.

S1: Well, I'm looking at like, 12 hours you get 3 people, and then 5, 5 [at 14 and 15 hours], you know, more people study more hours if they have a curfew.

S2: But, there's also more people.

S1: But, I mean it's like less people who don't [have a curfew]. You know, there's like 1 for 12 hours, 2 for 14. Do you know what I mean?

S2: Yeah.

S1 focused on a relatively small portion of the data. She apparently considered 12 – 15 hours as representing significant study time so that when she noticed that there were more students in this range with curfews than without, she concluded that those with curfews were studying more. Although it bothered S2 that there were more students overall with curfews, she could not offer a compelling argument for why this was important and in the end accepted S1's analysis.

We have now observed many students, (from 5th graders to college seniors) using this slicing technique across a range of problems. An interesting feature of this method of comparing groups is that it eliminates the problem of how to compare two groups with variable elements. Comparisons are made only at points where cases in each group have the same, or nearly the same, value. This may explain why students in Konold et

al. (1997) preferred two way tables and students in Hancock et al., (1992) preferred Venn Diagrams to other types of displays. Both of these representations allow students to clearly identify the case values on the dependent variable of interest. This makes it easy (compared to, say, box plots to form subgroups with identical values on that dependent variable so that they can then compare the frequencies of the cases in each comparison group.

In summary, despite the fact that instruction in statistics tends to focus on measures of center, many students do not use those measures when they would be particularly helpful — to make comparisons between groups composed of variable elements. We suggest that this pattern is symptomatic of students not having adopted an interpretation of averages as central tendencies (or propensities, as we referred to them in Konold et al., 1997). As we discuss below, there are a variety of interpretations that can be given to an average. But most of these provide no clear conceptual basis for comparing two groups.

Alternative Interpretations of Center

Not all data have a central tendency, as we have defined it above. We could compute the mean weight of an adult elephant, a Mazda, and a peanut, but there is no clear process being measured here which we can regard as having a central tendency. One might think that the mean weight of all the elephants in a particular zoo might be a central tendency. But without knowing more about how the elephants got there or their ages, it is questionable whether this mean tells us anything about a process with a central tendency. Quetelet described this distinction in terms of "true" means of distributions that followed the law of errors and "arithmetic" means that could be calculated for any assortment of values such as our olio above (see Porter, 1986, p. 107).

We describe below a number of other ways to think of averages, including viewing them as fair shares, data reducers, and typical values. Some of these interpretations are described in Strauss and Bichler (1988) as "properties" of the mean. Others are described by Mokros and Russell (1995) as "approaches," which they observed elementary and middle school students using. We consider an "interpretation" to be

the goal a person has in mind when he or she computes or uses an average. It is the answer a person might give to the question "Why did you compute the average of those values?"

Although we describe these interpretations below in terms of the mean, most of them apply equally well to other measures of center, such as the median. In Table 2 we have also listed beside each interpretation an illustrative problem context. Of course, any problem could be interpreted from a variety of perspectives. But we chose the examples in Table 2 because it seems to us that their authors intended them to elicit a particular interpretation.

Table 2. Example contexts for various interpretations of averages.

Interpretation	Example context
Data reduction	Ruth brought 5 pieces of candy, Yael brought 10 pieces, Nadav brought 20, and Ami brought 25. Can you tell me in one number how many pieces of candy each child brought? (From Strauss & Bichler, 1988)
Fair share	Ruth brought 5 pieces of candy, Yael brought 10 pieces, Nadav brought 20, and Ami brought 25. The children who brought many gave some to those who brought few until everyone had the same number of candies. How many candies did each girl end up with. (Adapted from Strauss & Bichler, 1988)
Typical value	The number of comments made by 8 students during a class period were 0, 5, 2, 22, 3, 2, 1, and 2. What were the typical number of comments made that day? (Adapted from Konold & Garfield, 1992.)
Algorithmic	What is the average of the following numbers? 10.2, 14.3, 9.7, 11,0, 12.6 (From Hardiman, Well & Pollatsek, 1984).
Formal relations	The mean age of 5 persons in a room is 30 years. A 36-year-old person walks in. What is now the mean age of the persons in the room? (From Moore, 1985, p. 196)
Central tendency	A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student were 6.2, 6.0, 6.0, 15.3, 6.1, 6.3, 6.2, 6.15, 6.2. What would you give as the best estimate of the actual weight of this object? (Adapted from Konold & Garfield, 1992.)

Data reduction. Here an average is seen as a way to boil down a set of numbers into one value. One of the students interviewed by Konold et al., (1997) gave the following rationale for why she would look at a mean or median in trying to describe the number of hours worked by students at her school.

> We could look at the mean of the hours they worked, or the median.... It would go through a lot to see what every, each person works. I mean, that's kind of a lot, but you could look at the mean.

The data need to be reduced because of their complexity, in particular because of the difficulty of holding in memory the individual values. The student above went on to say: "You could just go through every one... [but] you're not going to remember all that."

Fair share. The computation for the mean is probably first encountered in elementary school in the context of fair-share problems, with no reference to the result being a mean or average. Quantities distributed unevenly among several individuals are collected and then redistributed evenly among the individuals. The word "average," in fact, is derived from the Arabic '"awariyah," which translates as "goods damaged in shipping." According to Schwartzman (1994), the Italians and French borrowed this term to refer to the financial loss resulting from damaged goods. Later it came to specify the "portion of the loss borne by each of the many people who invested in the ship." Strauss and Bichler (1988) provided eleven problems as examples of tasks they used in their research, eight of which we regard as fair-share problems.

Typical value. An average is seen in this interpretation as some sort of typical score. This includes ideas related to the majority, mode, median, and midrange. This is perhaps the most frequently encountered interpretation seen in current pre-college curricula. Students are often given a data set and asked to determine a "representative" or "typical" value. There are, however, two different ways of thinking about a typical value. When educators pose this question to students, they are presumably thinking of a value that is representative of the entire group. But many students' responses suggest that they believe a typical value describes a characteristic of particular case, or set of

cases, in the distribution. Below is an example of this usage among 3rd graders who were discussing a display showing the heights of students in their class (from Konold and Higgens, in preparation). Referring to their heights, their teacher had asked them "How do we decide what's average?"

Phoebe: I think I'm taller than average because I notice that on the playground.

Brita: I was right. Sam is average, and I'm average too. We are the same.

Tiffany: I'm average too.

Katie: I'm not average. I'm shorter.

To claim that within a group of children Sam is of "typical" or "average" height is a statement about Sam, and not necessarily about the group as a whole.

The use of averages to describe particular individuals rather than groups is supported by common usage where we frequently speak of the "average student." We see this attribution in the phase "'typical' student or basketball player" as quoted above by Bright and Friel (1998), and it seems to be this notion of the mean that Gould (1996, p. 41) critiques when he claims that "no actual individual can stand for the category's deeper reality." Here, apparently even Gould got pulled into thinking that the individualistic interpretation of a mean as typical value is its "real" interpretation.

Algorithm. According to this interpretation, the meaning of an average is synonymous with the mathematical operations that produce it. This is one of the approaches used by students interviewed by Mokros and Russell (1995). It is an interpretation we find almost exclusively in school settings where students are asked to perform various mathematical operations with no other purpose than to demonstrate computational competence. The task of finding the average of a set of numbers out of context, as in the example in Table 2, seems intended only to test whether students know how to compute the mean. In this case, a perfectly acceptable answer to the question "Why did you compute the mean of those numbers?" is "Because I was asked to." This may seem so impoverished an interpretation that it should not be considered an interpretation at

all. However, we see students responding to problems presented in more meaningful contexts in ways that suggest that their view of an average is dominated by how it is computed, and thus it qualifies in our mind as an interpretation (Pollatsek, Lima & Well, 1981).

Formal relations. Many of the tasks students encounter in school explore their understanding of formal properties of averages. These include understanding how the mean of a set of numbers is related to the sum of those numbers, how it functions as a balance point, and how the sum of the deviation scores is zero. For the median, it would include understanding when changing the value of a score would affect the median, and when it would not. Interpretations related to formal relations would also include more qualitative ideas such as the fact that the mean lies somewhere within the range of the set and that it need not be a value in that set. We consider many of the properties of the mean that Strauss and Bichler (1988) explore to be interpretations involving formal relations.

Applying These Interpretations to the Problem of Group Comparison

Earlier in the paper we explored how the statistician's interpretation of central tendency provides a basis for using means in the NAEP example to compare groups. We claim that none of the alternative interpretations we describe above provides a clear rationale for why anyone should take the difference between the means of the two years seriously. Furthermore, compared to the interpretation of central tendency, other interpretations provide little conceptual basis for even comparing two groups using averages. It is one thing to summarize a group of data in some way using a statistic; it is quite another to take that statistic seriously enough as to use it to represent the *entire* group as one must when comparing the averages of two groups.

Contrasting the statistician's view of a central tendency with the data reduction view helps illustrate this point. The statistician regards a central tendency as providing information that is not directly available in the individual values: a central tendency gauges the "signal" that is masked in the individual values, a signal that becomes evident only in the aggregate. According to the data reduction interpretation, data are

distilled down to a single value out of a necessity to simplify. There is nothing in this interpretation which suggests that any new information emerges from this process; indeed a considerable loss of information seems to be the price paid for reducing complexity. On this logic it would seem that as a data set grows larger, the single-value summary becomes less representative of the group as increasingly more information is lost in the reduction process.

Nearly the same critique applies to the typical-value interpretation. But additional problems arise for those who think of a typical value as describing a characteristic of a particular individual. We suspect that the students Bright and Friel (1998) interviewed held this individualist view of typical value. If we asked them to describe exceptional cases instead of describing what was typical, they presumably would have pointed to a student or group of students who were particularly tall, for example. But we would not have expected them to use the height of the taller student to summarize the whole group. So why should they think of using the height of the typical student to do so? Again, our argument is that interpretations of averages based on typicality provide little or no support for using them to make a claim about the entire group.

The fair-share interpretation may provide some basis for using means to compare groups. One could think of the mean in the 1998 NAEP data as the reading score that all students sampled that year would have if reading ability were divided evenly among all the students sampled. Based on this reasoning, one might reasonably conclude that the 1998 group had a higher reading score than the 1994 group. But we would guess that many students would regard such reasoning skeptically unless in the real world situation it was physically possible to reallocate quantities — if, for example, we were talking about marbles or money.

The critique above is based largely on a logical analysis. However, the fact that many students who know how to compute various measures of center do not use them to compare groups supports this analysis. Whatever interpretations they give to averages including means and medians, these students do not view these statistics as sufficient for representing the group in a head-to-head comparison with another group.

Teaching Central Tendency

Our own thinking about how to teach students about measures of center has evolved over the years. Fifteen years ago, we were focused on how to help students develop a conceptual understanding along with their ability to compute means and medians. This often involved developing their understanding of the formal relations between means and their constituent elements, for example, coming to understand that the sum of the n values could be reconstituted from the mean, that the mean by itself gives no information about how the constituent values are distributed, that in small samples the mean can be heavily influenced by one outlier. We took it for granted that once students had developed this more complete set of understandings, they would know what in general to do with means and how to properly interpret them. Moreover, we thought we had bigger fish to fry, such as helping the university students we were teaching to understand the intuitions behind the Central Limit Theorem and the logic of statistical testing.

Observing that many of these same students did not think of using means when they were pursuing their own questions about group differences confused us (e.g., Konold et al., 1997). However, it eventually led us to reflect on interpretations that we give to measures of center, interpretations that, more than the purely formal aspects, instill confidence that averages are telling us something useful and alert us to instances in which their use is misleading or ridiculous.

Our reading of the historical development of the idea of central tendency suggests that this is not an idea that students will take to quickly. When early astronomers began computing means of observations, they were very cautious, if not suspicious, about whether and when this made sense. Before the middle of the eighteenth century they would never combine their own observations with those obtained from another astronomer because they were fearful that if they combined data that had anything but very small errors, that the process of averaging would multiply rather than reduce the effect of these errors (Stigler, 1986, p. 4). It was another hundred years before Quetelet began attributing central tendency to social and human phenomena, and this brought stiff rebukes from thinkers such as Auguste Comte who thought it ludicrous to believe

that we could rise above our ignorance of values of individual cases simply by averaging many of them (Stigler, 1986, p. 194). To Comte, statistics applied to social phenomena was computational mysticism.

Recently, statistics educators have made a serious effort to involve students early in the analysis of real (or at least realistic) data (e.g., Hogg, 1992; Lajoie, 1998). There are undoubtedly several good reasons for doing this. For one, real data make statistics more interesting. Our analysis makes clear a deeper rationale: that it is impossible to think about an average of disembodied numbers as having a central tendency. Students can practice with their calculator how to enter these values and extract a mean, but because there is no process specified, they certainly cannot regard that other than the numbers. That many of our former students have been exposed to decontextualized statistics may help explain why so many of them have had difficulty understanding what they mean.

But having some context for data we give students is not enough either. As we argued earlier, averages can be interpreted as central tendencies only in certain contexts. In selecting data for students, we should give careful consideration to the context and make sure there is a clear generating process. While it might be fine to pose to very young students the sorts of fair share problems we included in Table 2, we do not believe these are problems that we should be using when we want to encourage statistical thinking. Because it is relatively easy to motivate the computation of the average in fair share problems, some have recommended that this is the way to begin introducing means to students (Bakker, 1999). However, we claim that there is a critical difference between thinking of 2.5 as the number of cookies each student would get if they shared evenly the cookies they all brought vs. interpreting the 2.5 as the average number of cookies that they brought. And in this case, it is difficult to think about a coherent process for which the 2.5 would be a gauge.

As important as the type of data we give students are their reasons for analyzing them. Too often we still find students being asked to compute averages when the reason for doing so is unclear (Feldman, Konold & Coulter, 2000). On the other hand, situations that involve determining the best possible value from a set of measurements or the

problem of comparing groups motivates the need to develop group descriptions such as the mean (Konold & Higgins, in preparation). It is critical that students view situations they are investigating as important, interesting, or as having practical implications. But even when problems are well motivated, it is easy once students begin talking about data to divorce themselves from the context, to forget the meaning of the values and why the are looking at them. When we ask students to look at a distribution of values and simply describe its features —range, center, shape — it is easy for them to forget about the context and their questions, and focus on visual aspects only.

Our analysis suggests, additionally, that students may find it more or less difficult to think about averages as central tendencies depending on the general type of process they are investigating. Our intuition is that situations involving repeated measures are conceptually the easiest to think about as having central tendencies. In looking through many of the materials developed for elementary and middle school students, we find that repeated measures are rarely used, perhaps because they often lack pizzazz. However, it may be a fruitful place to introduce students to the concept.

References

American Psychological Association (1996). Task force on statistical inference initial report. http://www.apa.org/science/tfsi.html.

Bakker, A. (1999). Historical and didactical phenomenology of the mean values. In the Proceedings of the Conference on History and Epistemology in Mathematics Education in Belgium.

Bright, G. W., & Friel, S. N. (1998). Helping students interpret data. In Lajoie, S. P. (Ed.), Reflections on statistics: Learning, teaching, and assessment in grades K-12 (pp. 63-88). Mahwah, NJ: Lawrence Erlbaum.

Donahue, P. L., Voelkl, K. E., Campbell, J. R., & Mazzeo J. (1999). *NAEP 98 reading report card for the Nation and the States.* Document No. NCES 1999500. Washington, D.C.: National Center for Educational Statistics, U.S. Department of Education. http://165.224.221.98/pubsearch/pubsinfo.asp?pubid=1999500

Feldman, A., Konold, C., & Coulter, R., with Conroy, B., Hutchison, C., & London, N. (2000). *Network science, a decade later: The internet and classroom learning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Frick, R. W. (in press). Interpreting statistical testing: Processes, not populations and random sampling. *Behavior Research Methods, Instruments, & Computers.*

Gal, I., Rothschild, K., & Wagner, D. A. (1990). *Statistical concepts and statistical reasoning in school children: Convergence or divergence.* Paper presented at the annual meeting of American Educational Research Association. Boston.

Gould, S. J. (1996). *Full house.* New York: Harmony Books.

Hacking, I. (1990). *The taming of chance.* Cambridge: Cambridge University Press.
Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27(3),* 337-364.
Hardiman, P. T., Well, A. D., & Pollatsek, A. (1984). Usefulness of a balance model in understanding the mean. *Journal of Educational Psychology, 76(5),* 792-801.
Hogg, R. V. (1992). Towards lean and lively courses in statistics. In F. S. Gordon and S. P. Gordon (Eds.), *Statistics for the twenty- first century,* (MAA Notes, #26, pp. 3-13). Mathematical Association of America.
Konold, C., & Garfield, J. (1992). *Statistical Reasoning Assessment: Intuitive Thinking.* Unpublished Manuscript. Amherst: University of Massachusetts.
Konold, C., & Higgens, T. L. (in preparation). Working with data. In S. J. Russell, D. Schifter, & V. Bastable, *Developing Mathematical Ideas.* Parsippany, NJ: Dale Seymour Publications.
Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analyzing data: Research of critical barriers. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: 1996 Proceedings of the 1996 IASE Round Table Conference* (pp. 151-167). Voorburg, The Netherlands: International Statistical Institute.
Lajoie, S. P. (Ed.). (1998). *Reflections on statistics: Learning, teaching, and assessment in grades K-12.* Mahwah, NJ: Erlbaum.
Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20-39.
Moore, D. S. 1992. Teaching statistics as a respectable subject. In F. S. Gordon and S. P. Gordon, eds., *Statistics for the twenty-first century* (MAA Notes, #26). Washington, D.C.: Mathematical Association of America.
Moore, D. S. (1985). *Statistics: Concepts and Controversies* (Second Edition). New York: W. H. Freeman and Company.
Plackett, R. L. (1970). The principle of the arithmetic mean. In E. S. Pearson and M. G. Kendall (Eds.), *Studies in the history of statistics and probability* (pp. 121-126. London: Charles Griffen & Company.
Pollatsek, A., Lima, S., & Well, A. (1981). Concept or computation: Students' misconceptions of the mean. *Educational Studies in Mathematics, 12,* 191-204.
Porter, T. M. (1986). *The rise of statistical thinking 1820-1900.* Princeton: Princeton University Press.
Quetelet, M. A. (1842). A treatise on man and the development of his faculties. Edinburgh: William and Robert Chambers.
Schifter, D., Bastable, V., & Russell, S. J. (In preparation). Developing mathematical ideas: Collecting, representing, and analyzing data. Parsippany, NJ: Dale Seymour Publications.
Schwartzman, S. (1994). *The words of mathematics: An etymological dictionary of math terms used in English.* Washington, D.C.: Mathematical Association of America.
Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge: Harvard University Press.
Strauss, S. & Bichler, E. (1988). The development of children's concepts of the arithmetic average. *Journal for Research in Mathematics Education, 19 (1),* 64-80.
Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37,* 145-168.